



Working Paper 10 - 34
Statistics and Econometrics Series 18
September 2010

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

A Semiparametric State Space Model

André A. Monteiro¹

Abstract

This paper considers the problem of estimating a linear univariate Time Series State Space model for which the shape of the distribution of the observation noise is not specified a priori. Although somewhat challenging computationally, the simultaneous estimation of the parameters of the model and the unknown observation noise density is made feasible through a combination of Gaussian-sum Filtering and Smoothing algorithms and Kernel Density Estimation methods. The bottleneck in these calculations consists in avoiding the geometric increase, with time, of the number of simultaneous Kalman filter components. It is the aim of this paper to show that this can be achieved by the use of standard techniques from Cluster Analysis and unsupervised Classification. An empirical illustration of this new methodology is included; this consists in the application of a semiparametric version of the Local Level model to the analysis of the well-known river Nile data series.

Keywords: *Clustering, Gaussian-Sum, Kernel Methods, Signal Extraction, State Space Models*
JEL Classification: *C13, C14, C22*

¹ Universidad Carlos III de Madrid, C/ Madrid 126, Getafe 28903, Spain. Tel. +34634422955 Fax. +34916249849
email: andreavpbmonteiro@gmail.com

A Semiparametric State Space Model

André A. Monteiro*

Department of Statistics, University Carlos III of Madrid

First Version: September 2010

Abstract

This paper considers the problem of estimating a linear univariate Time Series State Space model for which the shape of the distribution of the observation noise is not specified a priori. Although somewhat challenging computationally, the simultaneous estimation of the parameters of the model and the unknown observation noise density is made feasible through a combination of Gaussian-sum Filtering and Smoothing algorithms and Kernel Density Estimation methods. The bottleneck in these calculations consists in avoiding the geometric increase, with time, of the number of simultaneous Kalman filter components. It is the aim of this paper to show that this can be achieved by the use of standard techniques from Cluster Analysis and unsupervised Classification. An empirical illustration of this new methodology is included; this consists in the application of a semiparametric version of the Local Level model to the analysis of the well-known river Nile data series.

Keywords: Clustering, Gaussian-Sum, Kernel Methods, Signal Extraction, State Space Models.

JEL classification codes: C13, C14, C22

AMS 2010 MSC codes: 62M10, 93E11, 93E14, 62G07

*Contact details: Department of Statistics and Econometrics, Universidad Carlos III de Madrid, C/ Madrid 126, Getafe 28903, Spain. Tel. +34634422955 Fax. +34916249849 email: andreavpbmonteiro@gmail.com

1 Introduction

The class of state space models provides a unified and flexible framework for modelling and describing a wide range of time series and other types of longitudinal data in a variety of disciplines. State space models are now routinely used in Time Series Analysis and other fields of Statistics where longitudinal data plays a role (e.g. Ansley and Kohn 1985; Durbin and Koopman 2001; Kitagawa 1987, 1989 and 1994; Kitagawa and Gersch 1996; and many others). However, the need to *fully* specify the underlying distributions in parametric state space models is a significant shortcoming of the framework. In fact, in many applications, there is no objective reason to assume a parametric form for the distributions involved, that is, if we exclude considerations related to computational convenience. Worse still, in several cases it is known that specific popular models cannot account for the statistical features of the data. This is, for example, the case with the basic Stochastic Volatility model, which assumes the normality of the return distribution conditional on the unobserved volatility (see, for example, Liesenfeld and Richard 2003).

Considering the pervasiveness of the use of (mostly linear and Gaussian) state space methods in current empirical work in Economics, Biostatistics, and several areas in Engineering, and taking into account that the linear and Gaussian assumptions are, at best, overly optimistic and, possibly, even inadequate, it would be desirable to investigate the possibility of relaxing the underlying, and somewhat restrictive, parametric assumptions. This paper represents one small step in that direction. Concentrating on the issue of the specification of the noise sequence corrupting the observation of the state variables, a univariate semiparametric state space model is introduced. This very simple model contains one single state variable driven (still) by Gaussian noise through a linear transition function. However, the density of the noise corrupting the observations is left *unspecified*, and is estimated directly from the observed time series. This model is potentially of interest for situations (i.e. applications) where it may be reasonable to assume that the underlying signal of interest in a time series results from the additive effect of many independent stochastic factors - such that a central limit theorem may be invoked. However, no strong, and arbitrary, distributional assumptions are made concerning the nature of the noise corrupting that signal.

In order to estimate both the underlying model parameters and the unknown observation noise density, this paper builds on standard results from the theories of generalized state space models and nonparametric density estimation. In particular, the well-established class of *Kernel methods* (for a textbook treatment see, for example, Silverman 1986) is employed for obtaining the observation noise density. In standard Kernel Density Estimation theory each observation

makes an elementary contribution to the overall density estimator. However, in the current setting, the ‘observations’ (i.e. the observation noise sequence) are in fact an *unknown* function of the data, and not the data itself. This fact adds an extra element of difficulty to the problem. However, by considering a linear form for both the state and observation equations and, most critically, employing a Gaussian kernel inside the observation noise density estimator, the observation noise sequence can be estimated from the data by standard linear filtering and smoothing methods. It is the aim of this paper to show that, not only is the problem of relaxing the parametric assumptions in state space models conceptually solvable, but also that the particular choices made here render the whole estimation task perfectly feasible, computationally. This can potentially open the way for developing a broader class of semiparametric and, eventually, even nonparametric state space methods for analyzing time series and more general longitudinal datasets. The main objective is combining both the current, and powerful, signal extraction algorithms associated with (both linear and nonlinear) state space models with the robust inference properties (with respect to the particular underlying distributions) of nonparametric density estimation methods.

This paper is the first development in a research agenda first proposed by the author in early 2006. Meanwhile, two papers have appeared that possess somewhat convergent objectives (Grillenzoni 2009; Rigat and Smith 2008). Grillenzoni (2009) combines the use of Kernel Density estimators with the classical (observation driven) ARMAX modelling methodology, in order to relax the parametric distributional assumptions of this class of Time Series models. It is well-known that ARMAX models can be expressed in state space form. Rigat and Smith (2008) propose a semiparametric state space modelling approach that is closer, and somewhat complementary, to the one taken in the current paper. Their approach essentially consists in exchanging the (parametric) state transition density with a semiparametric one obtained by combining a sequential nonparametric change-point testing procedure and a parametric transition equation. Each time a new observation is made, a nonparametric test for the occurrence of a change-point in the state vector is conducted. If such a change-point is deemed likely then a parametric transition prior is applied. This essentially means that the state transition density is assumed to have a point of mass, corresponding to the situation where there is no change in the state vector. In contrast, here, the focus is on the semiparametric treatment of the observation equation.

The remainder of this paper is organized as follows. In Section 2, the particular class of univariate semiparametric state space models considered in this paper is introduced. The associated filtering and smoothing algorithms are discussed, jointly with the procedure for constructing the Likelihood function. Section 3 discusses both the problem of the initialization of the filter-

ing recursions and the estimation of the unknown observation noise density in practice. Further implementation details of the semiparametric version of the Local Level model, used in the empirical section, are also discussed here. Section 4 presents and discusses the estimation results obtained in an illustrative empirical use of this model. The conclusions drawn from this study are presented in Section 5, jointly with some indications for future work.

2 The statistical model and related inference

As it was mentioned in the introduction, this paper represents a (small) step in the direction of developing a general class of semiparametric state space models for the analysis of multivariate time series and panel datasets. In order to illustrate the main ideas on how this problem can be approached (and to avoid getting unnecessarily hindered by the added complexities of having to handle multivariate time series), I focus on a simple univariate state space model with a single unobserved state variable. Nevertheless, the main ideas discussed here can be directly extended to a more general multivariate setting.

Statistical inference for the model introduced in this section includes estimating both the unknown model parameters and the density of the observation noise. The later is obtained as a kernel density estimate. In standard Kernel Density Estimation theory, one needs to choose both a kernel function and the so-called smoothing, or *bandwidth* parameter. The kernel function is, essentially, a bounded density function with finite second-order moments. While choice of the kernel function is relatively unimportant (in the sense that the density estimation results are fairly robust with respect to the use of different functional forms for the kernel, see Silverman 1986), the choice of the bandwidth parameter is critical. There is a very extensive Statistics literature dealing with the development of ‘objective’ (i.e. data-driven) methods for choosing the bandwidth. These can be broadly divided in two large groups; the first consists in methods based on the general principle of *cross-validation* (see for example Duin 1976; and Tanner and Wong 1984), while the second group comprehends the family of so-called ‘plug-in’ methods (e.g. Sheather and Jones 1991; Park and Marron 1990).

In the current setting, the bandwidth is simply an additional model (hyper-)parameter to be estimated from the data by Maximum Likelihood, thus requiring no special treatment.

Consider T observations of a univariate time series $(y_1, \dots, y_T) \equiv Y_T$. The corresponding simple state space model that is considered here is

$$\begin{cases} x_t = m_t x_{t-1} + \delta_t \eta_t & , \eta_t \sim N(0, 1) \\ y_t = \beta_t x_t + \epsilon_t & , \epsilon_t \sim g(\cdot), \end{cases} \quad (1)$$

where the time-invariant density g is left unspecified, and is to be approximated by its Kernel density estimate,

$$\hat{g}(\epsilon) = \frac{1}{(T+1)\lambda} \sum_{t=0}^T K\left(\frac{\epsilon - \epsilon_t}{\lambda}\right). \quad (2)$$

For the initial state x_0 a $N(a, \delta_0^2)$ distribution is assumed. Throughout this section it will be assumed that both a and δ_0^2 are known. In the empirical section this initial distribution will be treated as *diffuse*, that is, a will be set to an arbitrary value and the variance made arbitrarily large: $\delta_0^2 \uparrow \infty$ (following the general idea first introduced by Ansley and Kohn 1985).

The main problem with inference for generalized state space models consists in evaluating the integrals appearing in the formulas providing the general filtering, prediction and smoothing densities for the state variables (see Kitagawa 1987, formulas 2.2 - 2.5). It is well-known that when all the densities involved are Gaussian, these integrals have simple closed-form solutions that yield the famous Kalman filter and smoother algorithms (Kalman 1960). In order to be able to still employ these standard algorithms for filtering and smoothing in the current setting, a Gaussian kernel will be used. That is

$$K(d) = \frac{1}{\sqrt{2\pi}} \exp(-d^2/2). \quad (3)$$

This links this new model directly with the literature on Gaussian-sum based estimation of state space models, which started originally with Sorenson and Alspach (1971), and was later fully developed in Kitagawa (1989, 1994 and 1996) in the current Time Series context. In fact, under this choice for the smoothing kernel, model (1) becomes very close to the class of models analyzed in Kitagawa (1994). There are, however, some significant differences. In Kitagawa (1994) the mean of each individual Gaussian component, used to approximate the noise densities for both the observation and state equations, was restricted to zero. Therefore, estimation of the Gaussian-sum consisted in finding the variance of each component and the corresponding weight coefficients. Here, in contrast, each Gaussian component of \hat{g} is (ideally) centered on the unobserved individual disturbances ϵ_t . Furthermore, and following standard practice in kernel density estimation, all the different components of the Gaussian Kernel Density Estimator (KDE) have the same variance, and are given the same weight. Nevertheless, the link between model (1) and the class of models analyzed in Kitagawa (1994), means that the derivation of the filtering, prediction and smoothing recursions presented in this section directly parallels the procedures described in Sections 2.4 and 3 of that article. The filtering procedure, in particular, is obtained as a special case of the algorithm of Sorenson and Alspach (1971).

For the remainder of this section, assume that the observation disturbances $(\epsilon_0, \epsilon_1, \dots, \epsilon_T)$ are known exactly. In Section 3 we discuss a feasible approach for proceeding in practice, that is, when the observation disturbances are, in fact, unknown and must be estimated from the observed time series.

Let $\phi(d; \mu, \sigma^2)$ denote the probability density function of the Gaussian distribution with mean μ and variance σ^2 evaluated at the real number d . Writing the KDE (2) explicitly as a general Gaussian-sum

$$\hat{g}(\epsilon) = \sum_{j=0}^M \alpha_j \phi(\epsilon; c_j, W_j), \quad (4)$$

(where $M = T$, $c_j = \epsilon_j$, $W_j = \lambda^2$ and $\alpha_j = 1/(T+1)$) clearly shows that the semiparametric state space model, as defined by equations (1) and (2), using a Gaussian kernel (3) - and when the initial state density is a convex linear combination of Gaussian densities, falls under the scope of the Filtering algorithm of Sorenson and Alspach (1971). Accordingly, for any time-step $t = n$ the filtering (posterior) density $p(x_n|Y_n)$ has the general structure,

$$p(x_n|Y_n) = \sum_{l=0}^{M_n} \xi_n^{(l)} \phi\left(x_n; \mu_{n|n}^{(l)}, \sigma_{n|n}^{(l)2}\right), \quad (5)$$

(with $\sum_{l=0}^{M_n} \xi_n^{(l)} = 1$) meaning that the *one-step ahead state prediction* density is of the form,

$$p(x_{n+1}|Y_n) \propto \sum_{l=0}^{M_n} \xi_n^{(l)} \phi\left(x_{n+1}; \mu_{n+1|n}^{(l)}, \sigma_{n+1|n}^{(l)2}\right), \quad (6)$$

where

$$\begin{cases} \mu_{n+1|n}^{(l)} = m_{n+1} \mu_{n|n}^{(l)}, \\ \sigma_{n+1|n}^{(l)2} = m_{n+1}^2 \sigma_{n|n}^{(l)2} + \delta_{n+1}^2. \end{cases} \quad (7)$$

The corresponding predictive density for the next observation becomes,

$$p(y_{n+1}|Y_n) \propto \sum_{l=0}^{M_n} \sum_{j=0}^M \xi_n^{(l)} \alpha_j \phi\left(y_{n+1}; \hat{y}_{n+1|n}^{(l,j)}, \nu_{n+1|n}^{(l,j)}\right), \quad (8)$$

where

$$\begin{cases} \hat{y}_{n+1|n}^{(l,j)} = \beta_{n+1} \mu_{n+1|n}^{(l)} + c_j, \\ \nu_{n+1|n}^{(l,j)} = \beta_{n+1}^2 \sigma_{n+1|n}^{(l)2} + W_j. \end{cases} \quad (9)$$

The next filtering density, resulting from bringing in the information contained in a new observation y_{n+1} , is then,

$$p(x_{n+1}|Y_{n+1}) = \sum_{l=0}^{M_n} \sum_{j=0}^M \xi_{n+1}^{(l,j)} \phi\left(x_{n+1}; \mu_{n+1|n+1}^{(l,j)}, \sigma_{n+1|n+1}^{(l,j)2}\right), \quad (10)$$

with

$$\begin{cases} \mu_{n+1|n+1}^{(l,j)} = \mu_{n+1|n}^{(l)} + K_{n+1}^{(l,j)} \left(y_{n+1} - \hat{y}_{n+1|n}^{(l,j)} \right), \\ \sigma_{n+1|n+1}^{(l,j)2} = \left(1 - K_{n+1}^{(l,j)} \beta_{n+1} \right) \sigma_{n+1|n}^{(l)2}, \\ K_{n+1}^{(l,j)} = \sigma_{n+1|n}^{(l)2} \beta_{n+1} / \nu_{n+1|n}^{(l,j)}, \\ \xi_{n+1}^{(l,j)} \propto \xi_n^{(l)} \alpha_j \phi \left(y_{n+1}; \hat{y}_{n+1|n}^{(l,j)}, \nu_{n+1|n}^{(l,j)} \right). \end{cases} \quad (11)$$

In these filtering recursions, it should be noted that the so-called *innovation* $\left(y_{n+1} - \hat{y}_{n+1|n}^{(l,j)} \right)$, for component (l, j) , depends on the (unobserved) value of the observation disturbance ϵ_j , as seen in (9).

These recursions also imply that in the next time step $t = n + 1$ we have $M_{n+1} + 1 = (M + 1)(M_n + 1)$ terms in the Gaussian mixture (5). The ensuing ‘curse of dimensionality’ problem is the main shortcoming associated with the use of Gaussian-sum filtering. This explosion in the number of simultaneous Kalman filter components makes it virtually impossible, in practice, to exactly evaluate the Gaussian-sum filter for any realistically-sized time series. At least with currently, and commonly, available computer resources. Nevertheless, it was suggested in Kitagawa (1994), that this problem can be successfully overcome in practice by the use of a fixed number of Gaussian components at every time step $t = n$, following an idea of Harrison and Stevens (1976). In the current setting, the natural approach consists in constraining this number of Gaussian components $(M_n + 1)$ to be equal to $T + 1$. This is so, because $T + 1$ is the number of observations from which the KDE (2) of the unknown observation noise density is constructed. Nevertheless, this direct approach can still impose a heavy burden on limited computer resources, as discussed in Section 3.

In general terms, the problem consists in *approximating* (or summarizing) an arbitrary convex linear combination of a number N of Gaussian densities by another such Gaussian-sum - but one having a smaller number, say $M < N$, of individual Normal components. It is clear that this problem can be tackled using the techniques of *Cluster Analysis* and (unsupervised) *Classification*, for a textbook treatment of the subject see, for example, Anderberg (1973), Gordon (1999) or Gan et al. (2007). In fact the dimensionality-reduction problem at hand can be reformulated in terms of a clustering problem. Considering the N Gaussian components as our ‘observations,’ the aim is grouping these into M similarity clusters (or classes), and then summarizing each cluster by a single, representative, Gaussian component.

The suggestion in Kitagawa (1989 and 1994), consisted in applying a hierarchical single-link (binary) agglomerative clustering method, based on a symmetric dissimilarity measure between two Gaussian densities derived from the corresponding Kullback-Leibler information numbers. This dissimilarity measure is defined for a *pair* of Gaussian densities $\phi_i = (\mu_i, \sigma_i^2)$, $\phi_j = (\mu_j, \sigma_j^2)$

as,

$$D(\phi_i, \phi_j) = \frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} + \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} - 2. \quad (12)$$

For the current setup this would imply, after obtaining the $(T + 1)^2$ Gaussian components of the updated filtering density (10), *pooling* the two components that minimize this dissimilarity measure (i.e. the two most *similar* Gaussian components) into a single Gaussian component with the same overall mean,

$$\mu_{(i,j)} = \frac{\xi_i \mu_i + \xi_j \mu_j}{\xi_i + \xi_j},$$

and the same resulting variance

$$\sigma_{(i,j)}^2 = \frac{\xi_i [\sigma_i^2 + (\mu_{(i,j)} - \mu_i)^2] + \xi_j [\sigma_j^2 + (\mu_{(i,j)} - \mu_j)^2]}{\xi_i + \xi_j}.$$

And, obviously, the weight of the resulting component is given by the sum of the weights of the two original components $\xi_{(i,j)} = \xi_i + \xi_j$. This procedure would have to be repeated $(T + 1)T$ times until we are again left with only $T + 1$ Gaussian components. This is equivalent to sectioning the corresponding underlying binary *dendrogram* at the level that implies a partition of the original $(T + 1)^2$ Gaussian components into $T + 1$ similarity clusters (which is also the maximum split partition at this level: see for example Gordon 1999, section 3.4.1 and section 4.2.2). However, if this heavy computational procedure can still be feasibly implemented on a personal computer when the number of Gaussian components in the original (prior) noise densities is very small (for example just two, as in Kitagawa 1989 and 1994), the same no longer holds when there are many Gaussian components. This is, for example, the case in Section 4, where the number of Gaussian components in the observation noise equals 101 (i.e. the number of time-steps plus one). With limited computer resources, the total amount of calculations that have to be performed can lead to excessively large computer times making the whole approach unfeasible.

Remarkably, it turns out that for obtaining the same maximum split partition that results from applying the single-link agglomerative hierarchical clustering procedure described above, we can make use of a significantly faster partitioning algorithm. If we erase the $M - 1$ ‘heaviest’ edges in a *Minimum Spanning Tree (MST)* connecting all the N Gaussian components of the updated filtering density (10), considered as points of the (μ, σ^2) -plane, we obtain a maximum split partition into M similarity clusters (see, for example, Gordon 1999, section 3.4.1 or Anderberg 1973, section 6.4.2). In this analogy, the weight, or length, of each edge connecting two components (i.e the nodes of the graph) ϕ_i, ϕ_j is precisely the dissimilarity measure $D(\phi_i, \phi_j)$ given in (12).

The problem of finding the MST connecting the vertices in a given weighted graph is one of the oldest and most extensively-studied optimization problems in Computer Science (see Graham and Hell 1985). There are several computationally efficient algorithms available for solving this problem, in particular some whose time-complexity is linear or quasi-linear on the number of edges. In Section 4 Prim’s MST Algorithm (Prim 1957) is employed for performing these dimensionality reductions. However, other MST algorithms have been recently proposed which can be used for this purpose and are even faster (see Chazelle 2000; Karger et al. 1995). The bottom line is that it seems evident that it is possible to build an efficient computer implementation of the methodology discussed here, such that very large datasets (at least considerably larger than the time series of 100 yearly observations considered in Section 4) can be handled even with limited computer resources. Furthermore, as it is shown in Section 3, when available computer resources are indeed limited, it is straightforward to modify slightly the algorithms discussed here in order to significantly increase the speed of the associated computations. This can be done by using in place of the full KDE (2) a Gaussian-sum with a lower number of components (that is, lower than $T + 1$). As it was shown in Kitagawa (1994, section 5.2) it is possible to accurately approximate a given Gaussian-sum by another such density mixture having a lower number of components. Therefore, the significant gain in the speed of computations obtained by reducing the number of Gaussian components in the full KDE, should not be significantly penalized by a sizeable loss in accuracy in the resulting density estimate. At least if the reduced number of Gaussian components is ‘appropriately’ chosen, Section 3 discusses this point further.

Estimation, and inference, in the current setting is based on the Likelihood function associated with the model defined by equations (1), (2) and (3). From (8) the associated data density function is obtainable from the standard prediction error decomposition (see, for example, Durbin and Koopman 2001, section 2.10)

$$p(Y_T) = p(y_1) \prod_{t=2}^T p(y_t | Y_{t-1}). \quad (13)$$

Therefore, parameter estimation can be achieved by numerically maximizing the logarithm of the resulting Likelihood function.

Additionally, we are interested in obtaining *smoothed estimates* of the state variables x_n , $n = 1, \dots, T$, conditional on the entire time series Y_T . As it was shown in Kitagawa (1994) we can feasibly achieve this, when the noise distributions are given by Gaussian-sums, by the

application of the *Two-Filter* formula,

$$p(x_n|Y_T) = \frac{p(Y^n|x_n)p(x_n|Y_{n-1})}{p(Y^n|Y_{n-1})}, \quad (14)$$

where $Y^n = (y_n, \dots, y_T)$. The second factor in the numerator of (14) $p(x_n|Y_{n-1})$ was already obtained from the filtering algorithm, more specifically in the one-step ahead prediction step (6). The remaining element, the factor $p(Y^n|x_n)$ appearing in the numerator - which is taken as a density for x_n , must be carefully built through means of the following *Backward Filtering* algorithm.

We initialize the recursion with

$$p(Y^T|x_T) = p(y_T|x_T) \propto \sum_{l=0}^M \alpha_l \phi \left(x_T; \frac{y_T - c_l}{\beta_T}, \frac{W_l}{\beta_T^2} \right),$$

that is

$$\begin{cases} z_{T|T}^{(l)} = \frac{y_T - c_l}{\beta_T}, \\ U_{T|T}^{(l)} = \frac{W_l}{\beta_T^2}, \\ \gamma_T^{(l)} \propto \frac{\alpha_l}{|\beta_T|}. \end{cases}$$

The recursion step itself is as follows, assume that we have

$$p(Y^{n+1}|x_{n+1}) = \sum_{l=0}^{B_{n+1}} \gamma_{n+1}^{(l)} \phi \left(x_{n+1}; z_{n+1|n+1}^{(l)}, U_{n+1|n+1}^{(l)} \right),$$

then, from the elementary principle $p(Y^{n+1}|x_n) = \int p(Y^{n+1}|x_{n+1})p(x_{n+1}|x_n)dx_{n+1}$ we obtain the *one-step back prediction* density as

$$p(Y^{n+1}|x_n) \propto \sum_{l=0}^{B_{n+1}} \gamma_{n+1}^{(l)} \phi \left(x_n; z_{n+1|n}^{(l)}, U_{n+1|n}^{(l)} \right),$$

where

$$\begin{cases} z_{n+1|n}^{(l)} = z_{n+1|n+1}^{(l)} / m_{n+1}, \\ U_{n+1|n}^{(l)} = \left(\delta_{n+1}^2 + U_{n+1|n+1}^{(l)} \right) / m_{n+1}^2. \end{cases}$$

The previous backward filtering density is then given by

$$p(Y^n|x_n) = \sum_{j=0}^M \sum_{l=0}^{B_{n+1}} \gamma_n^{(l,j)} \phi \left(x_n; z_{n|n}^{(l,j)}, U_{n|n}^{(l,j)} \right),$$

where

$$\begin{cases} z_{n|n}^{(l,j)} = \left[z_{n+1|n}^{(l)} W_j + \beta_n U_{n+1|n}^{(l)} (y_n - c_j) \right] / J_n^{(l,j)}, \\ U_{n|n}^{(l,j)} = U_{n+1|n}^{(l)} W_j / J_n^{(l,j)}, \\ \gamma_n^{(l,j)} \propto \gamma_{n+1}^{(l)} \alpha_j \phi \left(y_n; \beta_n z_{n+1|n}^{(l)} + c_j, J_n^{(l,j)} \right), \\ J_n^{(l,j)} = \beta_n^2 U_{n+1|n}^{(l)} + W_j. \end{cases}$$

Here, because $B_n + 1 = (M + 1)(B_{n+1} + 1)$, it is also necessary to apply the clustering procedure previously described for constraining the number of Gaussian components to remain equal to a fixed number at all time-steps.

Finally, putting everything together we obtain the smoothing density

$$p(x_n | Y_T) = \sum_{l=0}^{M_{n-1}} \sum_{j=0}^{B_n} \omega_n^{(l,j)} \phi \left(x_n; \hat{x}_{n|T}^{(l,j)}, V_{n|T}^{(l,j)} \right)$$

with

$$\begin{cases} \hat{x}_{n|T}^{(l,j)} = \left(\mu_{n|n-1}^{(l)} U_{n|n}^{(j)} + z_{n|n}^{(j)} \sigma_{n|n-1}^{(l)2} \right) / L_n^{(l,j)}, \\ V_{n|T}^{(l,j)} = U_{n|n}^{(j)} \sigma_{n|n-1}^{(l)2} / L_n^{(l,j)}, \\ \omega_n^{(l,j)} \propto \gamma_n^{(j)} \xi_{n-1}^{(l)} \phi \left(z_{n|n}^{(j)}, \mu_{n|n-1}^{(l)}, L_n^{(l,j)} \right), \\ L_n^{(l,j)} = \sigma_{n|n-1}^{(l)2} + U_{n|n}^{(j)}. \end{cases}$$

Additionally, from the smoothing density for the state variables we can obtain indirectly the corresponding smoothing densities associated with the disturbance sequences $\{\eta_t\}$ and $\{\epsilon_t\}$.

3 Numerical implementation

In order to illustrate the algorithms described in the previous section, a particular case of the general model (1), known in the literature as the *Local Level* model, is applied to a widely studied time series. This is a data set consisting of a series of 100 readings of the annual flow volume of the river Nile taken at Aswan from 1871 to 1970. This series was first presented and analyzed in Cobb (1978), it has been subsequently modelled and analyzed by many other authors (e.g. Dümbgen 1991; Macneill et al. 1991; Balke 1993; Atkinson et al. 1997; Zeileis et al. 2003; and others). In particular, it was analyzed in Durbin and Koopman (2001, chapter 2) as a means to illustrate the empirical application of the fully parametric (Gaussian) Local Level model. The numerical computations in the next section were implemented using both the **0x** matrix programming language of Doornik (2002), version 5.10 and the **C++** language, making use of the routines contained in the **Boost** Libraries (in particular the **Graph** library of Siek et al. 2002).

The version of the Local Level model considered here is,

$$\begin{cases} x_n = x_{n-1} + \delta \eta_n & , \quad \eta_n \sim N(0, 1) \\ y_n = x_n + \epsilon_n & , \quad \epsilon_n \sim g(\cdot) \end{cases} \quad (15)$$

In this way, we can concentrate on the essential and innovative aspect of model (1), that is, the nonparametric estimation of the unknown density of the observation noise.

Throughout Section 2, it was assumed that the observation disturbances $\vec{\epsilon} = (\epsilon_0, \epsilon_1, \dots, \epsilon_T)$

were known (or, equivalently, that we had access to what is sometimes called a ‘oracle’ estimator for these unknown quantities). Obviously, in practice it is necessary to estimate the sequence $(\epsilon_0, \epsilon_1, \dots, \epsilon_T)$ along with the unknown constants δ and λ . This paper proposes a simple iterative procedure for estimating the unobserved observation disturbances and the unknown model parameters. At each step, the sequence of smoothed estimates $(\hat{\epsilon}_0, \hat{\epsilon}_1, \dots, \hat{\epsilon}_T)$ obtained in the previous step is used as an approximation to the true observation disturbances. Using this sequence as input, we apply first the filtering algorithm described in Section 2 in order to build the (approximate) model likelihood. This likelihood is then numerically maximized with respect to the model parameters. Using the resulting approximate ML estimates, the Gaussian-sum smoothing algorithm described in Section 2 is applied in order to obtain a new approximation to the true disturbances $(\epsilon_0, \epsilon_1, \dots, \epsilon_T)$. These iterations can be stopped when convergence, regarding the vector of observation disturbances, is achieved (according to some suitable criterium). This recursive procedure is initialized using the smoothed observation disturbances from the auxiliary standard Local Level model,

$$\begin{cases} x_n = x_{n-1} + \delta\eta_n & , \eta_n \sim N(0, 1) \\ y_n = x_n + \vartheta\epsilon_n & , \epsilon_n \sim N(0, 1), \end{cases} \quad (16)$$

which is analyzed, for example, in Durbin and Koopman (2001, Chapter 2). Furthermore, model (16) also provides adequate starting values (with $\hat{\lambda}^2 = \hat{\vartheta}^2 - (T+1)^{-1} \sum_{j=0}^T \hat{\epsilon}_j^2$) for the numerical optimization of the (approximate) Likelihood implied by the model defined by equations (1) and (2).

Some comments are in order regarding the reason why here it is assumed that the state sequence starts at $t = 0$ and not at $t = 1$. Clearly, it is necessary to have $E[\epsilon_t] = 0$, for all t , in order for model (1) to be identifiable. Therefore, it is necessary to ensure that the KDE $\hat{g}(\epsilon)$ has zero mean. A simple device for achieving this is considering that the (unobserved) sequence of states x_t starts at $t = 0$ instead of $t = 1$, that the observation y_0 is *missing*, and then defining $\epsilon_0 = -\sum_{n=1}^T \epsilon_n$.

The diffuse initialization of the Gaussian-sum filter, equations (6) through (11), is implemented as follows. Consider the conditional density of x_0 given the empty sequence Y_0 , clearly, $p(x_0|Y_0) = p(x_0) = \phi(x_0; a, \delta_0^2)$. From the one-step ahead state prediction relations (6) and (7), we get, $p(x_1|Y_0) = p(x_1) = \phi(x_1; \mu_1, \sigma_1^2)$, where

$$\begin{cases} \mu_1 = a, \\ \sigma_1^2 = \delta_0^2 + \delta^2. \end{cases}$$

The predictive density is given by $p(y_1|Y_0) = p(y_1) = \sum_{j=0}^M \alpha_j \phi(y_1; \hat{y}_1^{(j)}, \nu_1^{(j)})$, where $\hat{y}_1^{(j)} = \mu_1 + c_j$, and $\nu_1 = \sigma_1^2 + W_j$. Then the first filtering density is

$$p(x_1|Y_1) = \sum_{j=0}^M \xi_1^{(j)} \phi(x_1; \mu_{1|1}^{(j)}, \sigma_{1|1}^{(j)2}),$$

where, when we let $\delta_0^2 \uparrow \infty$, and work out the details from equations (8) through (11) we are able to conclude that

$$\begin{cases} \mu_{1|1}^{(j)} = y_1 - c_j, \\ \sigma_{1|1}^{(j)2} = W_j, \\ \xi_1^{(j)} = \alpha_j. \end{cases}$$

These relations provide the basis for initializing the filtering recursions, that is equations (6) through (11).

The Likelihood (13) implied by the model is dependent - however only through the first factor $p(y_1)$, on both the arbitrary value a and the unbounded variance δ_0^2 . If we consider instead the *diffuse Loglikelihood*

$$\begin{aligned} \log L_D &= \lim_{\delta_0^2 \uparrow \infty} (\log L + \tfrac{1}{2} \log \delta_0^2), \\ &= -\tfrac{1}{2} \log(2\pi) + \sum_{t=2}^T \log p(y_t|Y_{t-1}), \end{aligned}$$

the influence of both a and δ_0^2 is removed. Furthermore, the values of δ and λ that maximize $\log L$ also maximize $\log L_D$. Therefore, when the initial state density is diffuse we find the estimates of δ and λ by maximizing $\log L_D$ instead of $\log L$.

Due to the presence of the one-step ahead prediction density in the Two-Filter formula (14) the use of a diffuse density for the initial state x_0 means that here, the smoothing density $p(x_1|Y_T)$ also requires a special treatment. Working out the relevant details, it turns out that

$$\begin{cases} \hat{x}_{1|T}^{(l,j)} = z_{1|1}^{(j)}, \\ V_{1|T}^{(l,j)} = U_{1|1}^{(j)}, \\ \omega_1^{(l,j)} = \gamma_1^{(j)}. \end{cases}$$

Unfortunately, the amount of computer power required to implement the estimation procedure proposed in this paper surpasses what was available to the author at the time. This is the case even making use of the MST-based clustering-procedure for controlling the explosion of the number of Kalman Filter components. As an indication, in the early experiments, the time necessary for one single evaluation of the data density implied by the model used in the empirical section - when using the hierarchical single-link binary agglomerative clustering algorithm - was around 135 minutes. This was obtained on a PC with 0.5 GB of RAM, built around an Intel Pentium 4 CPU with a clock frequency of 3.0 GHz and running the **0x** matrix

programming language (Doornik, 2002), version 5.10, under Windows XP SP3. Obviously, that for larger Time Series we can expect even longer computer times. With the more efficient MST-based clustering procedure, using Prim’s MST algorithm (as implemented in the **Boost Graph Library**), one likelihood evaluation still requires almost 2 minutes. This was the case using a Dynamic Link Library (DLL), built with the **C++** language, and which calls the **Boost Graph Library**, the former being invoked from within an **0x** program.

Considering that the numerical maximization of the implied log-likelihood function was carried out using the Simulated Annealing-based global optimization procedure of Corana et al. (1987) - and which is implemented in the **0x** function **MaxSA** - that typically requires a very high number of function evaluations for identifying a global optimum (typically tens of thousands), it is easy to see that implementation of the direct approach would require more powerful computer hardware.

Therefore, in order to keep the whole computations at a level manageable with the above-mentioned personal computer, a modification of the original algorithm was introduced. As it was mentioned in Section 2, this modification consists in approximating the original KDE (2) by a distinct Gaussian-sum, one having a collapsed number of components $M + 1 < T + 1$. This can be done quite naturally using precisely the same dimensionality-reduction (clustering) procedure that is used, with each time-step, for controlling the growth in the number of Gaussian components.

The ensuing problem consists in choosing the number of components $M + 1 < T + 1$. In principle, it would be possible, at least conceptually, to choose an optimal number $M + 1$ of clusters according to a so-called ‘stopping rule’ applied to a complete set of hierarchically-nested partitions obtained from an agglomerative clustering algorithm - see for example Gordon (1999, Section 3.5). This basically requires defining a target ‘goodness’ criterium - computable for each possible *partition* of the $T + 1$ Gaussian components (from the original KDE) into $M + 1$ clusters. In the current context, a natural criterium could be specified as the Kullback-Leibler information number between the original KDE and the approximating Gaussian-sum - penalized by a strictly increasing function of M . Afterward a search for the value $M + 1$ which optimizes this goodness criterium would have to be conducted. However, computational considerations have also to be taken into account. Typically $T + 1$ will be a (somewhat) large number, therefore searching all the T possible values for $M + 1$ will require a fair amount of computer time.

A compromise solution consists in using some fixed reduction-rule, for example, we could directly define $M + 1 = \text{ceiling}(\sqrt{T + 1})$ or alternatively $M + 1 = \text{ceiling}[\log(T + 1)]$ (where $\text{ceiling}(x)$ denotes the smallest integer exceeding the real number x). This approach mimics

the current practice in introductory Descriptive Statistics when choosing the number of classes for grouping the observations of a continuous variable, for example when building an histogram or a frequency-table.

In Section 4 the number of collapsed components for approximating the KDE, in the empirical illustration of the methodology proposed here, was taken as $M + 1 = \text{ceiling}(\sqrt{T + 1})$, as $T = 100$ observations this leads to $M + 1 = 11$ Gaussian components. While this choice represents an approximation to the exact KDE, nevertheless it already enables the estimation of a much more flexible state space model, when compared with (16). One that already contains the main feature of the model defined by (1) and (2), that is, a *data-driven* specification for the density of the observation noise.

The filtering and smoothing recursions discussed in Section 2 as well as their initialization - as discussed here - remain perfectly valid. This is because (4) remains valid, but for the values of M , c_j and W_j . While the first, M , was already extensively discussed, the mean c_j and the variance W_j , of each Gaussian component, are obtained as the output of the dimensionality-reduction (clustering) procedure. With these modifications, and using also the combination of the C++ built DLL and an Ox program, one single evaluation of the Likelihood function typically takes less than 0.7 seconds, on the above-mentioned personal computer. Therefore, full estimation of the Local Level model (15) becomes perfectly feasible.

4 Empirical results

As mentioned in Section 3, in order to illustrate the use of the new methodology proposed in this paper, model (15) was fitted to a series of 100 yearly observations of the flow volume of the river Nile (measured in $10^8 m^3$) taken at Aswan from 1871 to 1970. This times series, which appeared first in Cobb (1978), is currently freely available for conducting Academic and Scientific research. It can be found on several servers on the internet (for example as part of the Time Series Data Library maintained by R.J. Hyndman¹). This time series is depicted in Figure 1.

<INSERT FIGURE 1 ABOUT HERE>

There are a few striking features: first some sort of structural change or level-shift seems to have taken place by the end of the nineteenth century and the beginning of the twentieth century. In fact, although, there are some large discharge volume readings during the twentieth century (for example 1916 and 1964), the yearly discharge volumes never returned to the pre-1898 lev-

¹<http://robjhyndman.com/TSDL>

els. Cobb (1978) provides evidence that indeed 1898 represented a change-point for this series. This conclusion is further corroborated by Dümbgen (1991), which provides a 95% confidence interval for the exact location of the change-point (this goes from 1896 till 1899). Also Zeileis et al. (2003), indicate 1898 as the most likely location of the single change-point for this series (after using the BIC criterium to determine the most likely number of change-points). In fact Cobb (1978) argues that this change was not just the result of the construction and beginning of operation of the first Aswan Dam (whose construction started in 1898, and started operating in 1902), but it was also due to an abrupt change in rainfall levels taking place by the end of the nineteenth century. Other competing explanations are proposed by other authors, in particular Macneill et al. (1991) argue that the Aswan Low Dam did not by itself cause the decrease in the flow volume of the Nile. However, they argue, the activities connected with the construction of the dam led to more accurate methods of measuring the flow volumes and hence to a downward correction of earlier upper-biased estimates.

In addition to the existence of (at least) one clear change, or variation, in the overall level of the series, there seem to be a few years with abnormally low discharge volume levels (relative to the surrounding years). In particular 1877, 1888 (and not 1887 as it reads, erroneously, on Atkinson et al. 1997, Section 6.3, on page 415) and 1913 seem good candidates for being labeled as outliers.

Finally, a general simple visual assessment of the series does give credibility to a decomposition of the type ‘trend plus noise.’

One word of warning to the reader must be left though: the fact that here, both the standard (fully Gaussian) and the new semiparametric Local Level models are fitted by Maximum Likelihood to this series, does not imply in any way that the author is trying to support the view that these simple models provide the *best* possible, global Statistical description of this series. It is obvious that that is not the case.²

Nevertheless, the Local Level model does provide a parsimonious, and perhaps sensible, tentative statistical description for this time series. It allows the extraction of an underlying trend, as well as the measurement of the variability of the readings with respect to this trend. These features enable this simple model to be, nevertheless, useful for answering meaningful questions related to this series; these can include (among others): the detection of level-shifts and abnormal years (outliers), as well as obtaining confidence bounds for future observations. However, here, the main objective of fitting the semiparametric Local Level model to this widely-studied

²For instance, it is not reasonable to assume that the flow volume of a river can take both arbitrarily large positive and negative values. However it is clear that models (15) and (16) allow unboundedly large values for the flow volume y_n (both negative and positive).

time series is to illustrate the use of this new methodology, and in particular contrasting it to the classic (fully parametric) structural time series approach.

As mentioned in Section 3, the first step for applying the new semiparametric Local Level model, consists in estimating the (classical) fully parametric version (16). This was done by Maximum Likelihood using the estimation and smoothing routines contained in the `0x` package `SsfPack`, version 2.2, of Koopman, Shephard and Doornik (1998). The corresponding results are shown in Table 1.

<INSERT TABLE 1 ABOUT HERE>

These results are essentially the same as those reported in Durbin and Koopman (2001, Table 2.1 on Section 2.10). Using the estimates of the observation disturbances, obtained by the fixed-interval smoother algorithm, associated with model (16) as input, we then apply the algorithms described in Section 2 and Section 3. As a simple stopping-rule, the iterations are repeated until the relative quadratic change in the resulting vector of observation disturbances $\vec{\epsilon}$ drops below 1%. For the dataset being analyzed this meant taking 5 iterations of the basic algorithm. The complete estimation results are shown in Table 2.

<INSERT TABLE 2 ABOUT HERE>

There are a few interesting features in these estimation results. First, the signal-to-noise ratio is rather uniform throughout the different iterations, and similar to the corresponding value obtained from the fully parametric model. This signal-to-noise ratio close to 10% indicates a high-level of variability (randomness) of the series around the estimated trend. Second, the variance of the state disturbances η increases consistently with the number of iterations, indicating that the trend obtained from the Semiparametric State Space Model (SPSSM) oscillates more than the one obtained from the Linear Gaussian State Space Model (LGSSM). This can be a hint that modelling this series as simply the result of an intervention (i.e. through means of a linear regression on a Heaviside step-function) plus white noise is, perhaps, an oversimplification (as done in Zeileis et al. 2003).

The extracted trend, resulting from the output of the fifth iteration, for the SPSSM is shown in Figure 2 together with what is obtained from the classical model.

<INSERT FIGURE 2 ABOUT HERE>

Although the trend extracted using the semiparametric model agrees, to a large extent, with the one obtained from the fully parametric model, there are some noticeable differences. As a consequence of the higher value for the variance of the state disturbances η we can see that the

trend from the SPSSM responds quicker to changes in the level of the series. A good example of this is the period that goes from 1894 till 1905. This is precisely the period where the steepest decrease in the average level of the yearly flow volume of the Nile river was felt. The trend obtained from the SPSSM shows a steeper slope than the equivalent trend from the LGSSM, and coming from a higher average ($1118.5 \cdot 10^8 m^3$ vs $1112.4 \cdot 10^8 m^3$) it actually reaches a lower end value ($799.3 \cdot 10^8 m^3$ vs $851 \cdot 10^8 m^3$ by 1905) - providing evidence that the decrease in the average flow volume level of the Nile may have been more severe than previously thought. It is interesting to note that in the period that goes from 1900 till 1908 the trend obtained from the LGSSM is far from the (local) average of the series, instead showing a clear upward-bias. This does not happen with the trend obtained from the SPSSM, this last one is very close to a local average of those few years. This overall pattern appears again in a few other places, in particular at the end of the series; in the period that goes from 1965 till 1970. More interesting still, in the period that goes from 1874 till 1880 the trend of the series, as obtained from the SPSSM, was reasonably higher than that obtained from the LGSSM. Inclusive, the highest value of the trend obtained with the SPSSM was achieved in 1878. This means that, despite an upward movement in the period from 1888 till 1893 (note that the period from 1890 till 1898 was an abnormal one with systematically above-average yearly flow volumes), the average level of the flow volume of the Nile was already decreasing by the early 1880's. This finding may provide supporting evidence for the view that the reason behind the decrease in the average flow volume of the Nile was not just the construction and beginning of operation of the Old Aswan Dam.

In order to get a feeling for the convergence behaviour of the SPSSM algorithm with respect to the extracted trend, Figure 3 plots all the 6 extracted trends (i.e. including the one obtained from the LGSSM).

<INSERT FIGURE 3 ABOUT HERE>

The behaviour is largely what could be expected, with a mostly uniform transition between the trend obtained with the LGSSM and the one derived from the fifth iteration of the SPSSM algorithm. Nevertheless, it is interesting to note that the final step-length vector from the Simulated Annealing optimization algorithm, reveals that the δ^2 parameter is actually harder to estimate than the squared bandwidth λ^2 . This may be a consequence of the small signal-to-noise ratio associated with the series.

Looking again at Table 2, and regarding the bandwidth parameter, the behaviour across the 5 iterations is not as regular. As we can see in Figures 5 through 9, in the first two iterations the estimate of the posterior density of the observation disturbances seems to be, clearly, under-

smoothed, showing excessive variability - a consequence of the relatively small estimated values for the bandwidth parameter λ . In contrast, in the third iteration, the bandwidth increased dramatically, leading to a density with only two local maxima (still far from a Gaussian density). Finally, in iterations 4 and 5 the bandwidth parameter seems to stabilize, leading to a stable estimated density for the observation disturbances; see Figure 8 and Figure 9. The general features of this density in the last two iterations are fairly identical. They both possess 4 prominent local maxima, one of these seems to be associated with the abnormally low flow volume years - a peak located around $-350 \cdot 10^8 m^3$. While the remaining are centered around $-75 \cdot 10^8 m^3$, $40 \cdot 10^8 m^3$ and $175 \cdot 10^8 m^3$. These estimated densities are strongly asymmetric. As a benchmark we also present, in Figure 4, a kernel density estimate applied to the (re-scaled) smoothed observation disturbances from the LGSSM, where the bandwidth parameter was automatically chosen using least squares cross-validation. Figure 4 also shows, superimposed, the plot of the Gaussian density with the same mean and standard deviation as the sample. Although not formally correct, this picture is, nevertheless, interesting as the KDE already shows two common features that also appear in the 5 next plots, corresponding to the semiparametric model. First, a heavier-than-Normal long left-tail that extends below $-400 \cdot 10^8 m^3$, and second, an excess of mass (with respect to the Normal density) concentrated between $200 \cdot 10^8 m^3$ and $300 \cdot 10^8 m^3$. Furthermore, there is a clear resemblance between Figure 4 and Figure 7. Nevertheless, the oversmoothing imposed by the cross-validation scheme does a good job in visually ‘Normalizing’ the smoothed observation disturbances from the LGSSM.

<INSERT FIGURE 4 ABOUT HERE>

<INSERT FIGURE 5 ABOUT HERE>

<INSERT FIGURE 6 ABOUT HERE>

<INSERT FIGURE 7 ABOUT HERE>

<INSERT FIGURE 8 ABOUT HERE>

<INSERT FIGURE 9 ABOUT HERE>

Finally, we look at the residual autocorrelation function (ACF) and partial autocorrelation function (PACF) of the smoothed observation disturbances. If the model is adequate for this

time series then the smoothed observation disturbances should resemble a pure (non-Gaussian) Noise process. Therefore both the sample ACF and PACF should not be significantly different from zero at all lags. Figure 10 presents these sample statistics (with the 95% confidence bands superimposed) for the smoothed observation disturbances obtained both from the LGSSM (16) and the output of the fifth iteration of the SPSSM algorithm.

<INSERT FIGURE 10 ABOUT HERE>

No clear differences between the two models are apparent from these sample statistics. All autocorrelations and partial-autocorrelations are non-significantly different from zero (perhaps with the exception of the PACF for the LGSSM at lag 10 - but this is still non-significant as we should expect 2 correlation values outside the confidence bands). Therefore, no further dynamics seem to exist in the estimated observation disturbances. As a final piece of evidence that the observation disturbances implied by the local level model (as applied to the present series) are, probably, not Gaussian, Figure 11 shows the Normal probability plots (usually known as QQ-plots) applied to each sequence of observation disturbances.

<INSERT FIGURE 11 ABOUT HERE>

The sequence of smoothed disturbances closer to Gaussianity, the one obtained from the LGSSM, still shows considerable deviations particularly below $-180 \cdot 10^8 m^3$ and above $160 \cdot 10^8 m^3$.

To finalize, I look into which observations can be likely labelled as outliers. By looking at Figure 9 it is reasonable to label as outliers any years for which the corresponding smoothed observation residuals fall below $-300 \cdot 10^8 m^3$ or are ‘close’ to $300 \cdot 10^8 m^3$. It turns out that there are only two observations falling under the first case (1877 and 1913) and none satisfying the second condition. Therefore 1888 cannot be reasonably considered an outlier, under the current modelling approach.

5 Conclusions and directions for further work

This paper has introduced a feasible algorithm for estimating a semiparametric univariate state space model. This was done with the purpose of illustrating the possibility of relaxing (at least some of) the parametric assumptions on which state space models are usually built. The main ideas behind this algorithm have the potential for being extended to larger dimensions. An empirical illustration of this new semiparametric dynamic modelling approach was included. It was, therefore, shown that the computations associated with this model could be implemented

in practice, even with limited computer resources. Using point estimates obtained from the new semiparametric state space model, a few new features were, potentially, uncovered in a well-known and extensively studied time series.

The new model introduced can potentially open the way for the development of new semi-parametric classes of state space models. It is, probably, not an overstatement to say that the potential for these new classes of models to uncover previously hidden patterns in longitudinal datasets is considerable. The technical challenges associated with the full development of these new classes of state space models may be also non-negligible. First, the computational implementation of these models seems to be slightly intricate (at least compared with classical fully Gaussian equivalents). Second, it is still necessary to extend to these new models the full set of inference tools currently available for fully parametric state space models. Therefore, a new line of research in state space methodology can, potentially, be open for future work.

References

- Anderberg, M.R. (1973). *Cluster analysis for applications*. Academic Press, New York.
- Ansley, C.F. and R. Kohn (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions, *Annals of Statistics*, **13**, pp. 1286 - 1316.
- Atkinson, A.C., Koopman, S.J. and N. Shephard (1997). Detecting shocks: outliers and breaks in time series, *Journal of Econometrics*, **80**, pp. 387 - 422.
- Balke, N.S. (1993). Detecting level shifts in time series, *Journal of Business and Economic Statistics*, **11**, pp. 81 - 92.
- Chazelle, B. (2000). A minimum spanning tree algorithm with inverse-Ackermann type complexity, *Journal of the ACM*, **47**(6), pp. 1028 - 1047.
- Cobb, G.W. (1978). The problem of the Nile: conditional solution to a change point problem, *Biometrika*, **65**, pp. 243 - 251.
- Corana, A., Marchesi, M., Martini, C. and S. Ridella (1987). Minimizing multimodal functions of continuous variables with the Simulated Annealing algorithm, *ACM Transactions on Mathematical Software*, **13**, pp. 262 - 280.
- Doornik, J.A. (2002). *Object-oriented matrix programming using Ox. 3rd ed.* Timberlake Consultants Press, London and Oxford. www.nuff.ox.ac.uk/Users/Doornik.
- Duin, R.W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Transactions on Computers*, **C25**, pp. 1175 - 1179.

- Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators, *Annals of Statistics*, **19**, pp. 1471 - 1495.
- Durbin, J. and S.J. Koopman, (2001). *Time series analysis by state space methods*. Oxford University Press, Oxford.
- Gan, G., Ma, C. and J. Wu (2007). *Data clustering theory, algorithms, and applications*. ASA-SIAM, Philadelphia.
- Gordon, A.D. (1999). *Classification*. Chapman & Hall, New York.
- Graham, R.L. and P. Hell (1985). On the history of the minimum spanning tree problem, *IEEE Annals of the History of Computing*, **7**(1), pp. 43 - 57.
- Grillenzoni, C. (2009). Kernel likelihood inference for time series, *Scandinavian Journal of Statistics*, **36**, pp. 127 - 140.
- Harrison, P.J. and C.F. Stevens (1976). Bayesian forecasting (with discussion), *Journal of the Royal Statistical Society Series B*, **38**, pp. 205 - 247.
- Kalman, R.E. (1960). A new approach to linear filtering and prediction problems, *Transactions of ASME, Journal of Basic Engineering*, **82D**, pp. 35 - 45.
- Karger, D.R., Klein, P.N. and R.E. Tarjan (1995). A randomized linear-time algorithm to find minimum spanning trees, *Journal of the ACM*, **42**(2), pp. 321 - 328.
- Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series, *Journal of the American Statistical Association*, **82**, pp. 1032 - 1041.
- Kitagawa, G. (1989). Non-Gaussian seasonal adjustment, *Computers & Mathematics with Applications*, **18**, pp. 503 - 514.
- Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother, *Annals of the Institute of Statistical Mathematics*, **46**(4), pp. 605 - 623.
- Kitagawa, G. and W. Gersch (1996). *Smoothness priors analysis of time series*. Springer-Verlag, New York.
- Koopman, S.J., Shephard, N. and J. Doornik, (1998). Statistical algorithms for models in state space form using SsfPack 2.2. *Econometrics Journal*, **1**, pp. 1-55.
- Liesenfeld, R. and J.F. Richard, (2003). Univariate and multivariate stochastic volatility models: estimation and diagnostics. *Journal of Empirical Finance*, **10**, pp. 505 - 531.
- Macneil, I.B., Tang, S.M. and V.K. Jandhyala (1991). A search for the Nile's change-points. *Environmetrics*, **2**(3), pp. 341 - 375.

- Park, B.U. and J.S. Marron (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, pp. 66 - 72.
- Prim, R.C. (1957). Shortest Connection Networks and Some Generalizations. *Bell System Technical Journal*, **36**, pp. 1389 - 1401.
- Rigat, F. and J.Q. Smith (2008). Semi-parametric dynamic time series modelling with application to detecting neural dynamics, *CRiSM Working Paper, Department of Statistics - University of Warwick*.
- Sheather, S.J. and M.C Jones, (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society Series B*, **53**(3), pp. 683 - 690.
- Siek, J.G, Lee, L.Q. and A. Lumsdaine (2002). *The Boost graph library: user guide and reference manual*. Addison-Wesley Professional, Reading MA.
- Silverman, B.W. (1986). *Density estimation for statistical and data analysis*. Chapman & Hall, New York.
- Sorenson, H.W. and D.L Alspach, (1971). Recursive Bayesian estimation using Gaussian sums, *Automatica*, **7**, pp. 465 - 479.
- Tanner, M.A. and W.H Wong, (1984). Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis, *Journal of the American Statistical Association*, **79**, pp. 174 - 182.
- Zeileis, A., Kleiber, C., Krämer, W. and K. Hornik (2003). Testing and dating of structural changes in practice, *Computational Statistics & Data Analysis*, **44**, pp. 109 - 123.

Table 1: Estimation Results Gaussian Local Level Model

This table contains the Maximum Likelihood estimates for the parameters of the auxiliary Gaussian Local Level model (equation (16) on page 11) - Iteration zero of the SPSSM algorithm. Remark: the LogLikelihood value presented corresponds to the concentrated diffuse LogLikelihood.

δ^2	ϑ^2	Signal-to-noise ratio	LogLikelihood
1469.2	15098	9.7311%	-632.546

Table 2: Estimation Results Semiparametric Local Level Model

This table contains Maximum Likelihood estimates for the parameters of the semiparametric Local Level model. Starting from the smoothed estimates of the observation disturbances from the auxiliary Gaussian Local Level model (equation (16) on page 11), the iterative procedure described in Section 3 was repeated until the relative quadratic percentage change in the vector of observation disturbances $\vec{\epsilon}$ dropped below 1%. Remark: the LogLikelihood values presented correspond to the diffuse LogLikelihood.

Iteration	δ^2	λ^2	Signal-to-noise ratio	LogLikelihood	% Quadratic Change in $\vec{\epsilon}$
1	1203.6	120.61	9.4824%	-630.42	13.679%
2	1420.7	113.92	9.8291%	-630.07	4.0087%
3	1539.8	1644.5	9.3759%	-631.93	6.7997%
4	1706.7	584.98	9.6217%	-632.45	1.9951%
5	2333.5	753.24	9.9104%	-632.07	0.320%

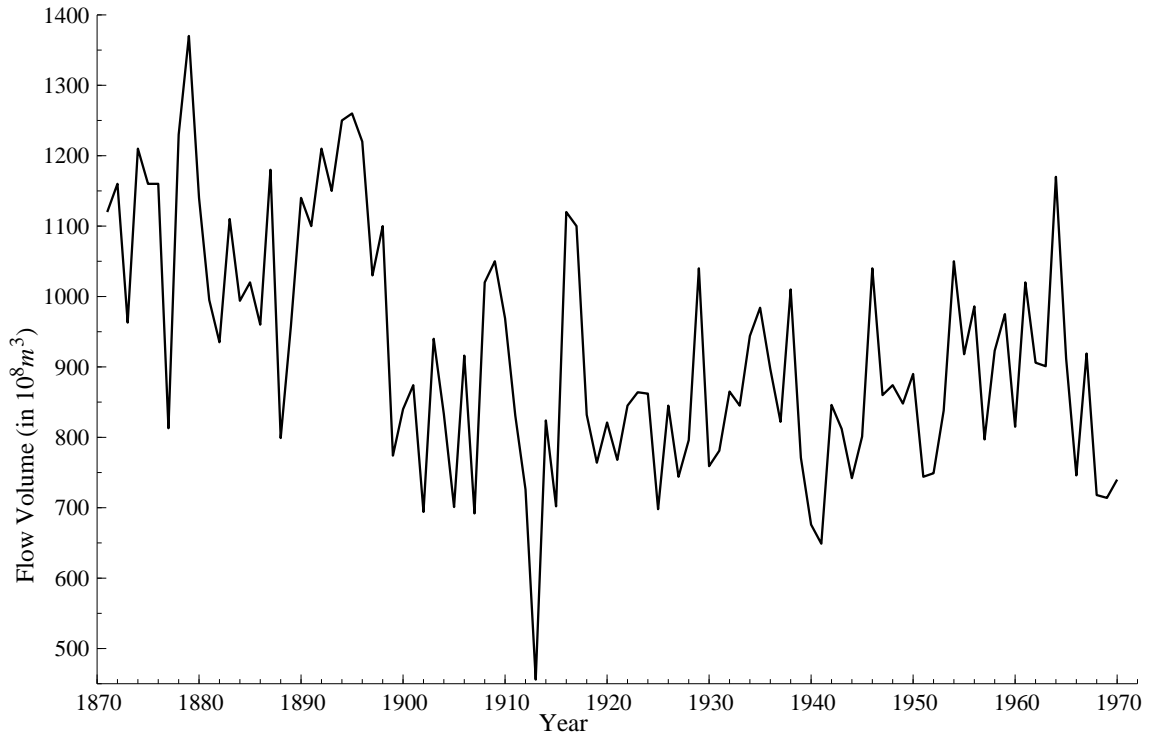


Figure 1: River Nile data

This graphic shows 100 years of observations of the flow volume of the river Nile taken at Aswan from 1871 till 1970.

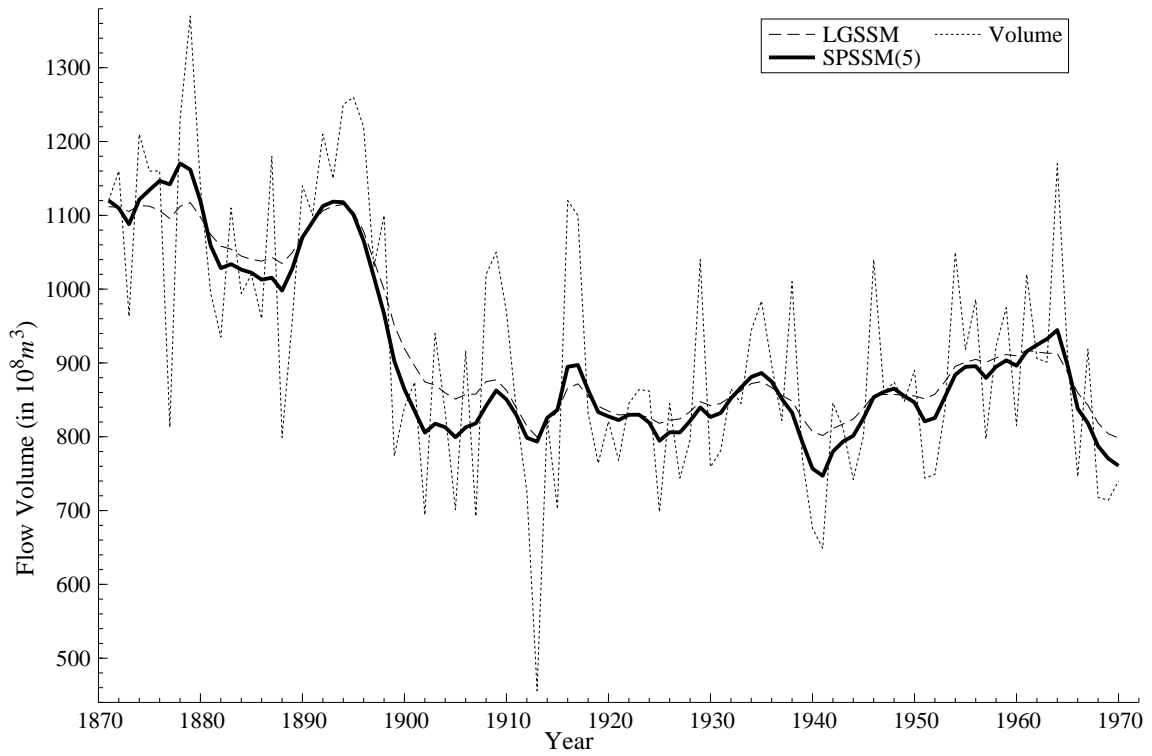


Figure 2: Extracted Trends - Gaussian vs Semiparametric models

This graphic shows the trends for the river Nile data, extracted using both the Gaussian and the Semiparametric (output from the fifth iteration) Local Level models.

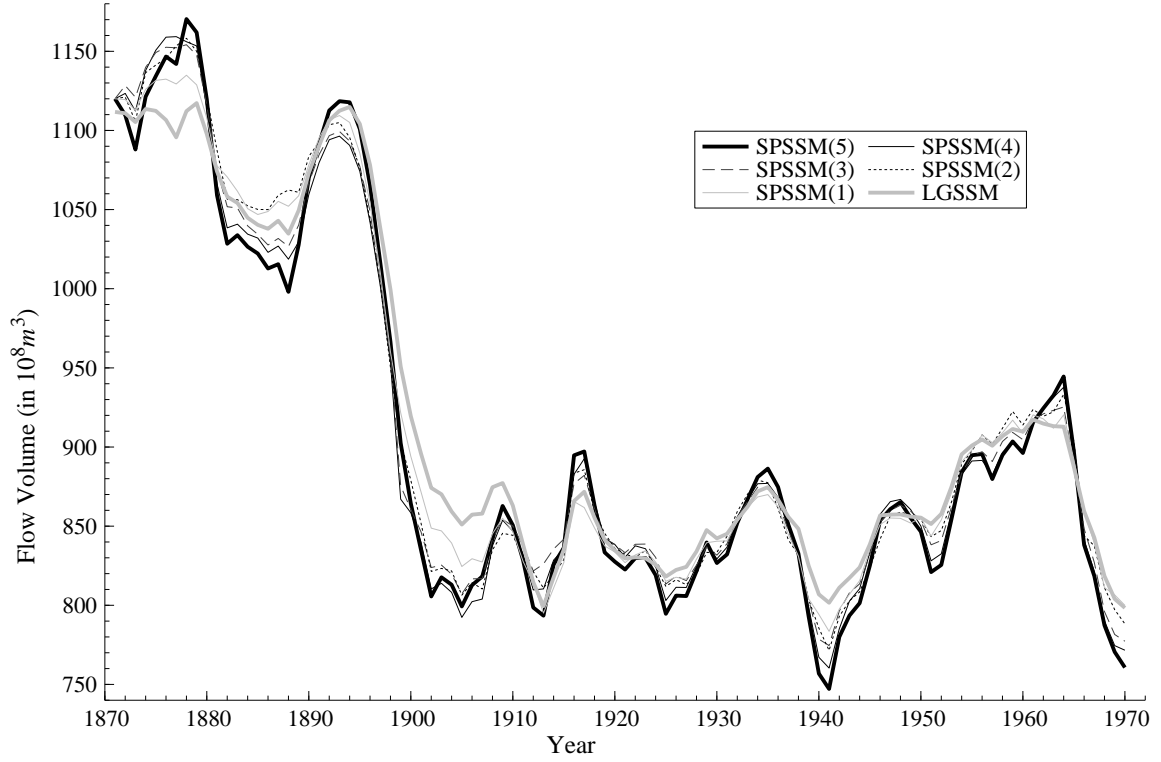


Figure 3: Extracted Trends - Gaussian vs Semiparametric models

This graphic shows a comparison of the trends extracted in each iteration of the procedure for estimating the Semiparametric Local Level model (see Section 3). The trend extracted using the fully parametric (classical) Gaussian Local Level model is further superimposed.

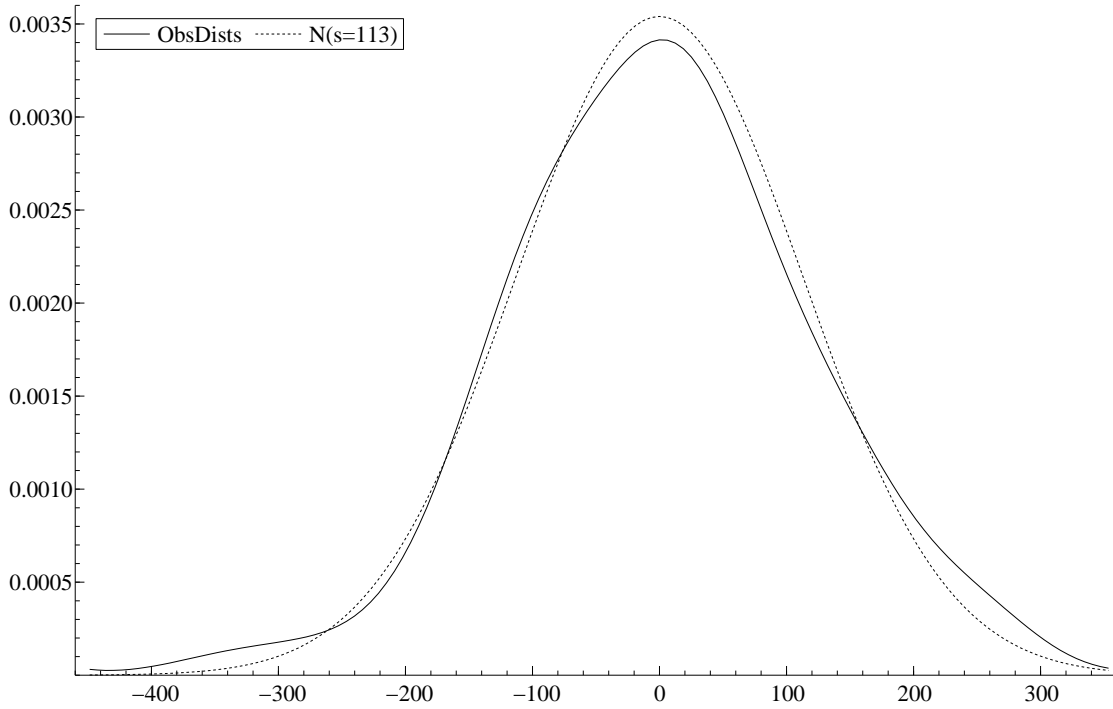


Figure 4: Estimated Density of the Observation Disturbances - LGSSM

A Kernel Density estimate was computed over the extracted smoothed observation disturbances from the Gaussian Local Level model. This acts as an informal assessment of the validity of the Gaussian assumption for the distribution of the observation disturbances in the Local Level model fitted to the Nile data.

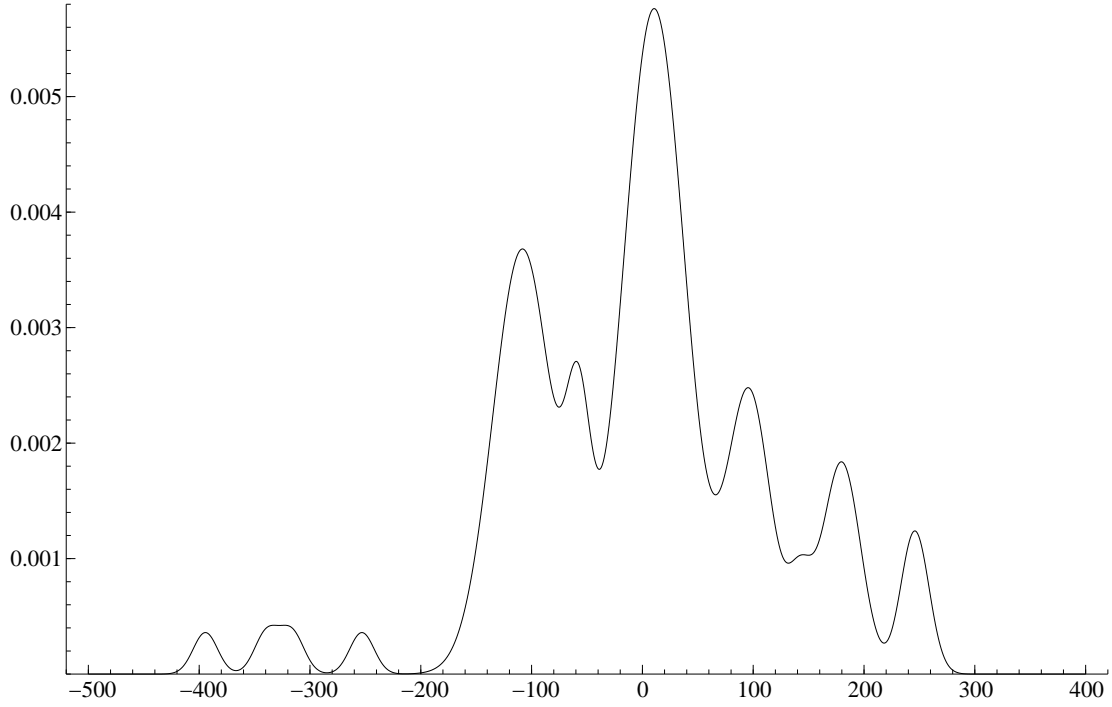


Figure 5: Estimated Density of the Observation Disturbances - 1st Iteration

This is the estimated density for the observation disturbances $g(\epsilon)$ obtained from the first iteration of the procedure for estimating the Semiparametric Local Level model (see Section 3).

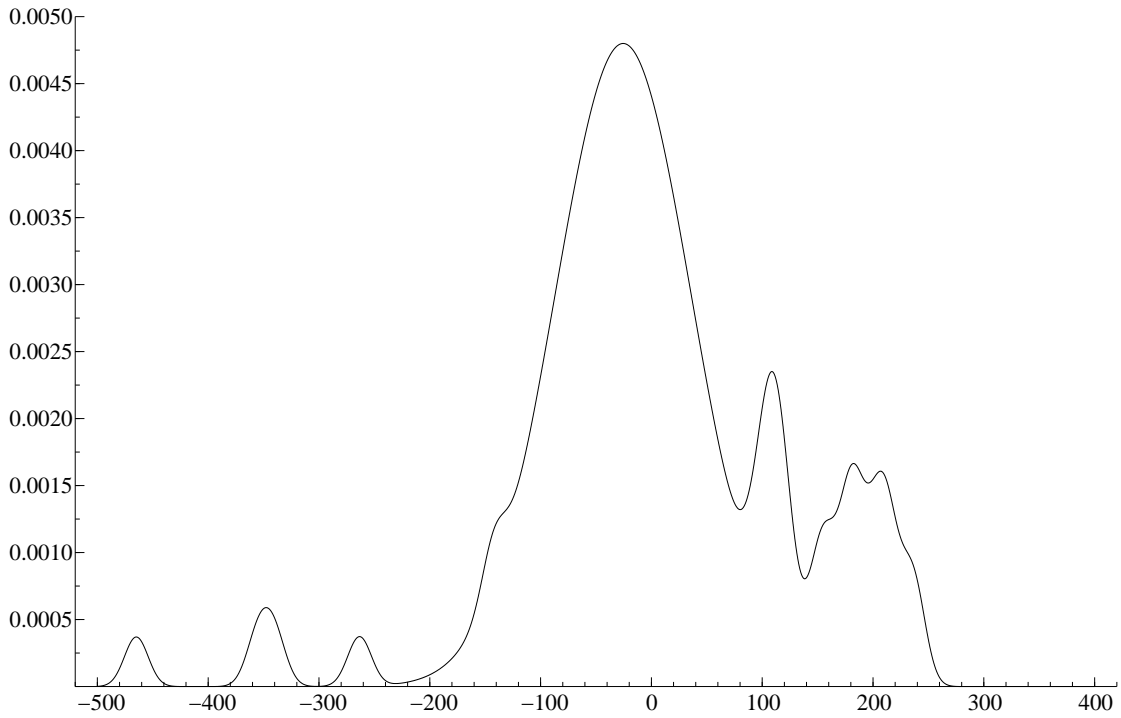


Figure 6: Estimated Density of the Observation Disturbances - 2nd Iteration

This is the estimated density for the observation disturbances $g(\epsilon)$ obtained from the second iteration of the procedure for estimating the Semiparametric Local Level model (see Section 3).

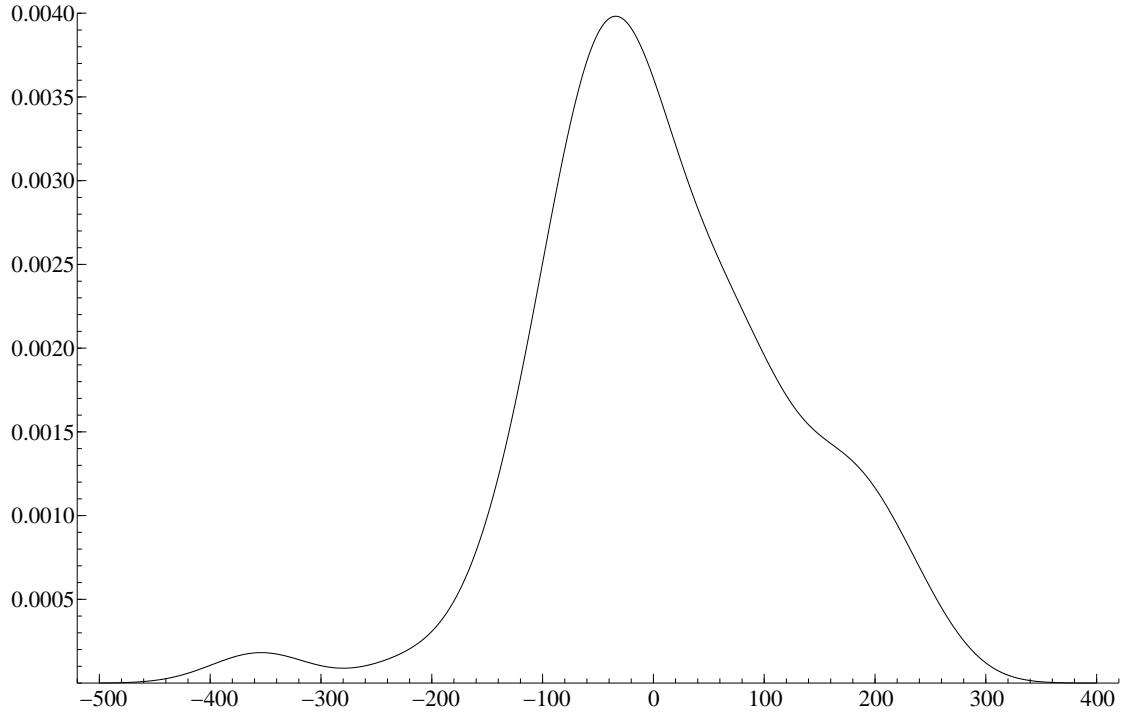


Figure 7: Estimated Density of the Observation Disturbances - 3rd Iteration

This is the estimated density for the observation disturbances $g(\epsilon)$ obtained from the third iteration of the procedure for estimating the Semiparametric Local Level model (see Section 3).



Figure 8: Estimated Density of the Observation Disturbances - 4th Iteration

This is the estimated density for the observation disturbances $g(\epsilon)$ obtained from the fourth iteration of the procedure for estimating the Semiparametric Local Level model (see Section 3).

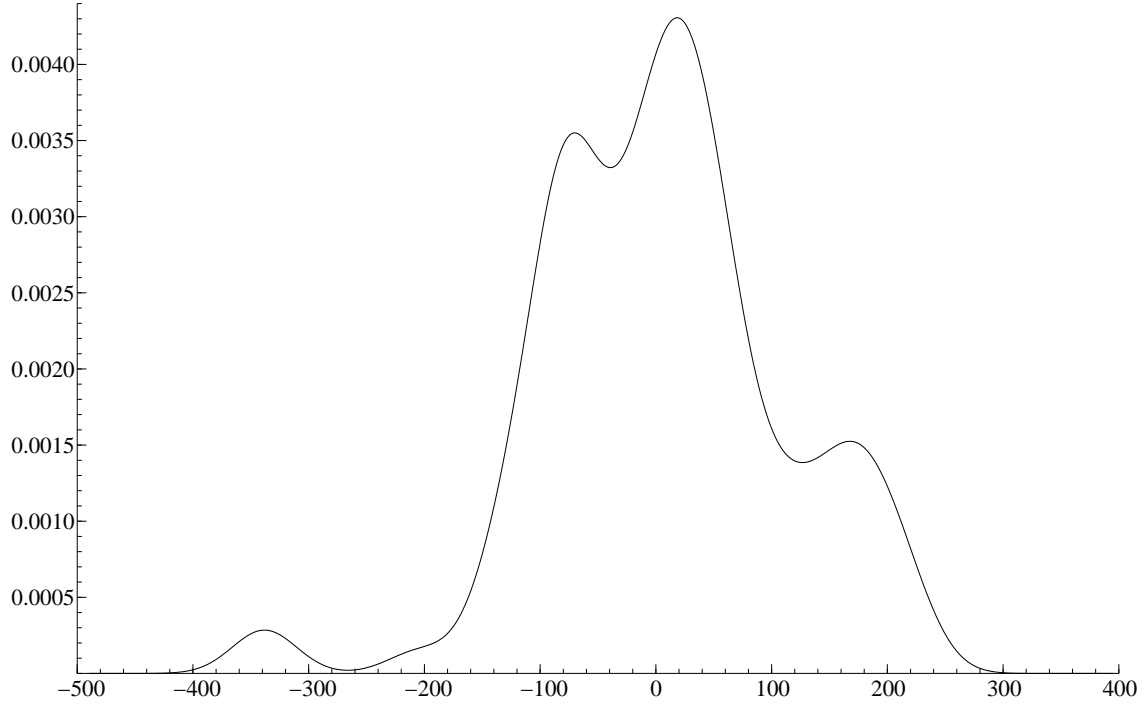


Figure 9: Estimated Density of the Observation Disturbances - 5th Iteration

This is the estimated density for the observation disturbances $g(\epsilon)$ obtained from the fifth and last iteration of the procedure for estimating the Semiparametric Local Level model (see Section 3).

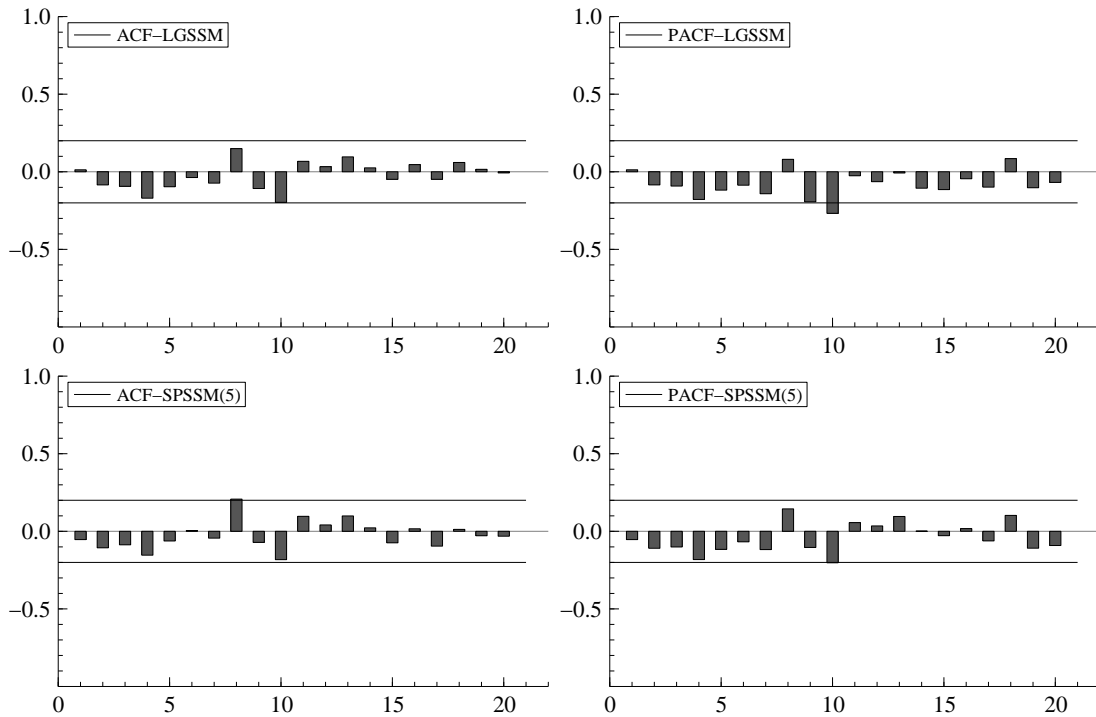


Figure 10: ACF and PACF of the Observation Disturbances

These are the estimated ACF and PACF for the smoothed observation disturbances obtained both from the standard Gaussian and the semiparametric Local Level (from the output of the fifth iteration) models). Also shown are the 95% confidence bands

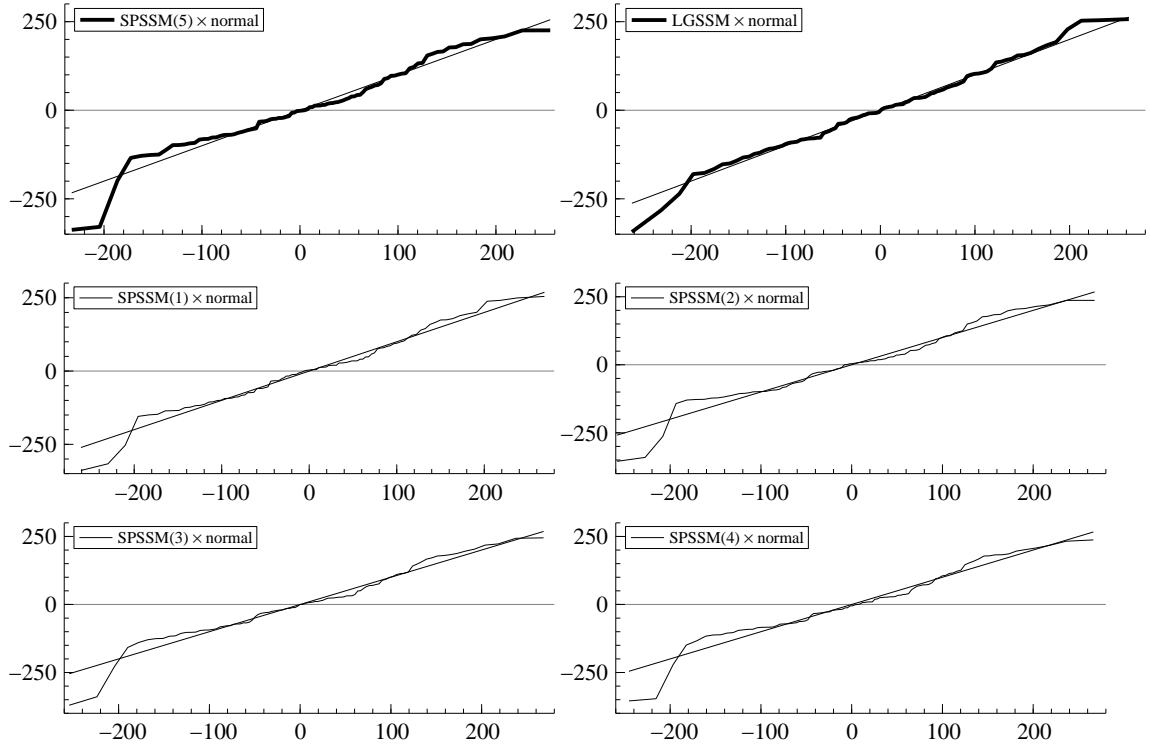


Figure 11: Normal QQ Plots of the smoothed Observation Disturbances

These are the Normal probability QQ-plots for the smoothed observation disturbances obtained both from the standard Gaussian and the semiparametric Local Level models (from the output of each iteration).